UPTEC X 03 010
FEB 2003

ISSN 1401-2138

ANNA SVENSSON

Hybridisation procedure and data analysis of a custom made microarray including 500 genes: a quality affirmation

Master’s degree project

Molecular Biotechnology Programme
Uppsala University School of Engineering

UPTEC X 03 010 | **Date of issue 2003-02**

Author

Anna Svensson

Title (English)

Hybridisation procedure and data analysis of a custom made microarray including 500 genes: a quality affirmation

Title (Swedish)

Abstract

Microarrays are powerful tools enabling the study of the expression level of a great number of genes simultaneously. The custom made microarray Myochip 1.0 provides a validated tool for the study of gene expression in human muscle biopsy material. To assure quality of raw data, dye-label concentration should be measured (i.e. in the Nanodrop spectrophotometer), minimal 10 pmol CyDye incorporation per sample is recommended. Normalisation of raw data is necessary to correct for systematic variation of ratios, print-tip group lowess normalisation appearing as a good choice. After normalisation, two well-working statistical methods to determine differentially expressed genes are SAM and Bayes, both showing high concordance with our own alternative "Eva". The data procedure has been tested and works well on biological data, predicting significant gene expression profile changes in response to strength training.

Keywords

cDNA microarray, differential gene expression, dye label quantity, normalisation, strength training

Supervisors

Prof. Eva Jansson

Institution of Medical Laboratory Sciences and Technology

Examiner

Ass. prof. Carl-Johan Sundberg

Institution of Medical Laboratory Sciences and Technology

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **29**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Hybridisation procedure and data analysis of a custom made microarray including 500 genes: a quality affirmation

Anna Svensson

Sammanfattning

Microarrays (s.k. genchip) är kraftfulla verktyg med vars hjälp man kan studera uttrycket hos ett stort antal gener i en och samma analys. Ett genchip består av korta DNA sekvenser som representerar olika gener, vilka placerats på ett substrat av glas. Den relativa skillnaden mellan två mRNA-prov undersöks genom att mäta hur väl dessa basparar till sekvenserna på genchipet. Genchip genererar stora mängder data, vilka måste bearbetas i flera steg. Först utförs en normalisering, för att avlägsna systematiska skillnader och påvisa biologiska skillnader mer tydligt. Därefter behöver man ett statistiskt verktyg för att hitta gener med förändrat uttryck.

Innan proven får baspara till genchipet märks respektive RNA-prov in med två olika fluorescerande färger. Hur väl denna inmärkning har lyckats bör mätas för att försäkra sig om pålitliga rådata. Vid första steget i den efterföljande dataanalysen fungerar normalisering med lokal linjär regression. För att sedan avgöra vilka gener som har ett signifikant ändrat uttryck kan man utnyttja någon form av modifierat t-test, t ex SAM och Bayes.

Ovanstående dataanalys har testats och visat sig fungera väl på biologiska data, en undersökning av hur geners uttryck förändras efter tre veckors styrketräning. Kunskapen om hur genuttrycket i human muskel påverkas av styrketräning är fortfarande mycket bristfällig. Förhoppningsvis kan nyttjandet av nya molekylärbiologiska metoder, som t ex genchip, leda till en ökad förståelse inom detta område.

Examensarbete 20 p i civilingenjörsprogrammet Molekylär bioteknik

Uppsala universitet februari 2003

Contents

1. BACKGROUND

1.1 Microarray technology

A *gene* consists of a segment of DNA which encodes a particular *protein*, the ultimate expression of genetic information. A *deoxyribonucleic acid* or *DNA* molecule is a double-stranded polymer composed of four basic molecular units called nucleotides. Each *nucleotide* comprises a phosphate group, a deoxyribose sugar, and *four nitrogen bases*. The four different bases found in DNA are adenine (A), guanine (G), cytosine (C), and thymine (T). The two nucleotide chains are held together by hydrogen bonds between nitrogen bases, with base pairing between G and C, and A and T respectively. The expression of genetic information stored in DNA is a two-stage process: (i) *transcription*, during which DNA is transcribed into *messenger ribonucleic acid* or *mRNA*, a single stranded complementary copy of the base sequence in the DNA molecule, with the base uracil (U) replacing thymine; (ii) *translation*, during which mRNA's nucleotide triplets are translated to *amino acids* specified by the *genetic code*. There are twenty different amino acids building up the proteins of the cell (1).

Microarrays are powerful tools enabling the study of the expression levels of thousands of genes simultaneously. Gene expression is analysed at the transcription stage, *i.e.* on mRNA level. Although regulation of protein synthesis in a cell can take place at any level in the process from DNA to protein, mRNA levels may sensitively reflect the type and state of the cell. Microarrays make use of DNA molecules property of *complementary base-pairing*. *Hybridisation* refers to the annealing of nucleic acid strands from different sources according to base-pairing rules. To utilise the hybridisation property of DNA, *complementary DNA* or *cDNA* is obtained from mRNA by reverse transcription (2).

cDNA microarrays are composed of individual DNA sequences, spotted on a high-density substrate of glass or nylon, a so-called *chip*. The relative difference between two RNA samples may be assessed by monitoring the differential hybridisation of the two samples to the sequences in the spots on the array. The samples, or *targets*, are reverse-transcribed into cDNA, labelled using different fluorescent dyes (e.g. a red- fluorescent dye Cy5 and a green-fluorescent dye Cy3), then mixed and hybridised with the spotted DNA sequences or *probes* (2,3).

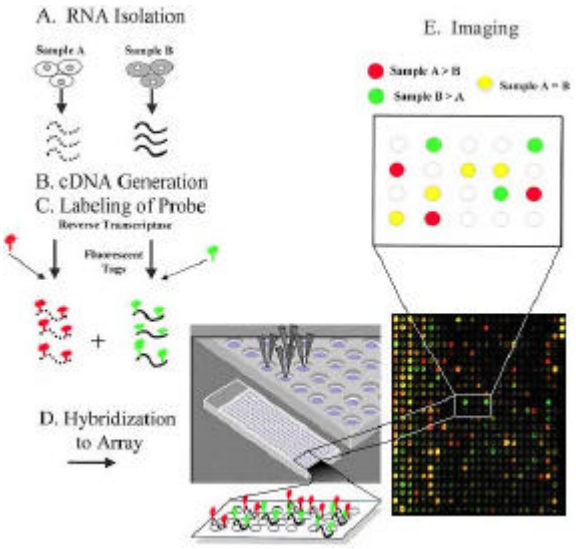

Fig 1. A cDNA microarray experiment, from RNA isolation to image analysis.
(The illustration was used with permission from Barry, R. FAO http://www.fao.org/DOCREP/003/x6884e03.htm, 1 Sep. 2002)

After this competitive hybridisation, the slides are imaged using a *scanner* and fluorescence measurements are made separately for each dye at each spot on the array. The ratio of the fluorescence intensity for each spot is indicative of the relative abundance of the corresponding DNA sequence in the two nucleic acid samples (2). The different steps in the microarray experiment procedure are illustrated in Fig. 1.

The microarray-technique has a wide range of applications including comparison of expression profiles after various drug treatments, classification of tumour cells, learning which genes are used in different cell types, studying gene expression during development (2) or: investigation of how expression profile change after a short period of strength training.

1.1.1 cDNA labelling

Various labelling systems label differently. The first step in the labelling procedure involves reverse-transcription from mRNA to cDNA. Here it is of great importance to use an efficient enzyme, which may reduce the amount of RNA required per reaction by more than a tenfold. When working with limited quantities of RNA it is of course enormously advantageous to make it possible to, for instance, use 2 instead of 20µg RNA (4).

Handling of the CyDyes also affects the labelling efficiency. One should always minimize the exposure of the dyes to all light sources by wrapping tubes in aluminium foil, if possible turn off the light in the lab and not store diluted pouches of dye longer than necessary (4).

It is also essential to have a high degree of *evenly* labelled targets. This can be achieved with an *indirect* labelling procedure; during cDNA synthesis, aminoallyl-modified dUTPs are first incorporated into the cDNA molecule. Next, in the coupling step, fluorescent dyes react with the modified dUTPs (Fig.2) (5).

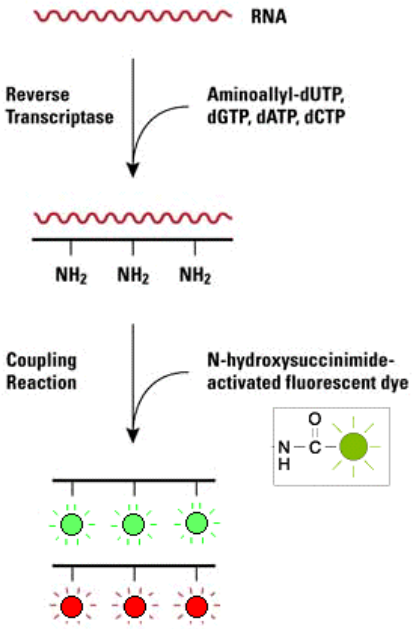

Fig 2. cDNA labelling. The indirect labelling procedure
(The illustration was used with permission from Hogarth, P.
ClonTech, Inc. http://www.clontech.com/archive/JAN02/Powerscript.shtml, 1 Sep 2002.)

A disadvantage with indirect labelling is that it is quite time consuming. An alternative is to use *direct* labelling, where first-strand cDNA is generated with dye-dNTP conjugate in one step. Because of their large size, dNTP conjugates are not efficiently incorporated for some dyes – particularly Cy5- potentially resulting in dye-biases (2,6).

Whatever labelling method chosen, the amount of CyDye in the samples should be measured before hybridisation. Knowing how much labelled material has been made will help in setting up correct and more reproducible hybridisation reactions and you do not risk wasting expensive microarray slides with substandard targets (4). This is most easily carried out by a spectrophotometer, measuring the absorbance at 550 nm for Cy3 and 650 nm for Cy5. The *Nanodrop™ND-1000 Spectrophotometer* allows measurements with 1µl sample volume and can detect Cy3 and Cy5 at concentrations as low as 0.1 pmol/µl (Fig. 3) (7).

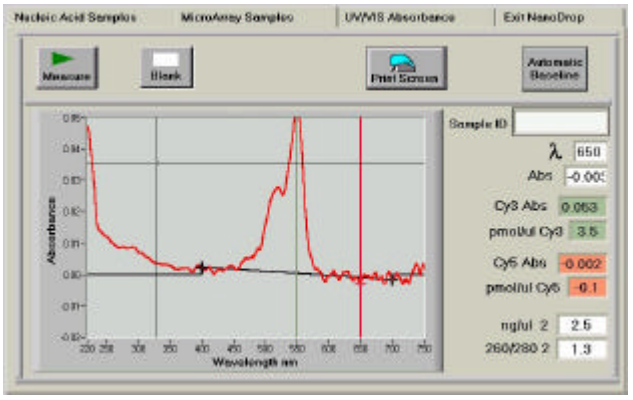

Fig 3. Dye label measurements Picture from the software connected to the Nanodrop™ND-1000 Spectrophotometer. On the screen, the green vertical line represents the peak wavelength position for Cy3, and the red line corresponds to the Cy5 ditto. This is a case of higher concentration of Cy3
(The illustration was used with permission from Stewart, J.
NanoDrop, Inc. http://www.clontech.com/archive/JAN02/Powerscript.shtml, 1 sep 2002.)

1.1.2 Scanning and data extraction

In the first step of image analysis, the hybridised arrays are imaged using a scanner. GenePix® 4000B uses a dual laser to scan the microarray at two wavelengths simultaneously, 532nm for detecting Cy3 and 635nm for Cy 5. The user adjusts the PMT (photo multiplier tubes) until the brightest spots are just below saturation ($2^{16} = 65\,500$), thus increasing the sensitivity of the image analysis for the weaker spots. Experiments with scanning a slide at varying PMT levels, however suggests that this has a negligible effect on the log-ratios and the ranking of genes (8, 9).

The red and green fluorescence intensities are already highly processed data. There are many alternatives for storing the output from a microarray experiment, but storing of the raw image files retains maximum information, allowing the use of different image extraction and quality metrices to be used subsequently (2).

Image processing is required to extract measures of transcript abundance for each gene spotted on the array from the laser scan images. The software GenePix Pro 4.0 is used for finding the spots and quantifying the signal intensities. Spots (or *features*) are grouped into rows and columns to form a *block*. Blocks are themselves grouped into rows and columns to form a template of the array. This template is then automatically aligned with the features on

the image (Fig. 4). The user must however manually check each spot and when necessary change their size or location. The user can also decide to flag odd looking features as "bad" (8).

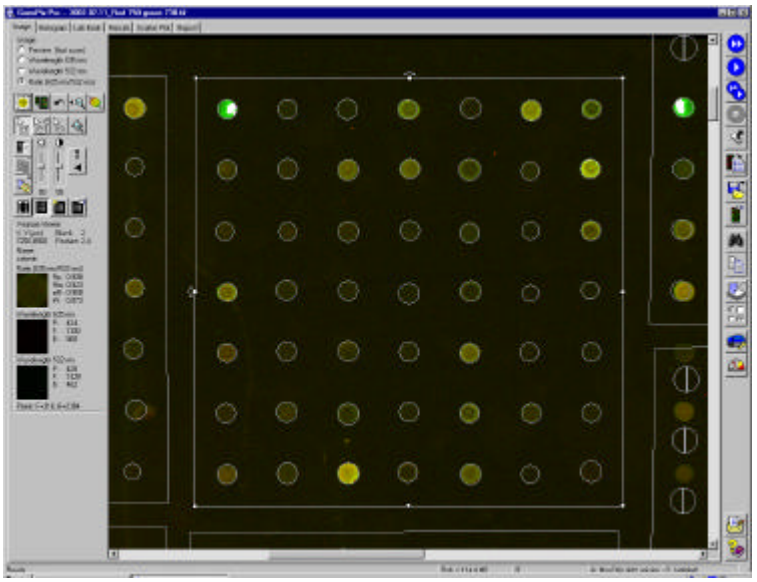

FIG 4. Data extraction software The GenePix Pro 4.0 user interface

1.1.3 Data displays

Microarray data is often logged for a number of reasons, *i.e.* the variation of logged intensities and ratios of intensities is less dependant on absolute magnitude, it makes normalisation additive and taking logs evens out skew distributions (10,11). Using unlogged data means that genes that are up-regulated by a factor 2 have an expression ratio of 2, whereas those down-regulated by the same factor have an expression ratio of (-0.5), resulting in all the down-regulated genes "squashed" between 1 and 0. By contrast, logarithms, which treat numbers and their reciprocals symmetrically, also treat expression ratios symmetrically. Up-regulated by a factor of 2 has $\log_2$(ratio) of 1, whereas a down-regulated gene by a factor of two has a $\log_2$(ratio) of –1. Genes expressed at a constant level (with ratios of 1) has $\log_2$(ratio) of 0 (12).

Expression data can be displayed by plotting the log intensity $\log_2 R$ in the red channel vs. the log intensity $\log_2 G$ in the green channel. This might however make interesting features of the data hard to see. An alternative choice is to plot log intensity ratio $M= \log_2(R/G)$ vs. the mean intensity $A= \log_2\sqrt{(R*G)}$ $(=0,5(\log_2 R+ \log_2 G))$, which facilitates the identifying of spot artefacts and detecting intensity dependent patterns in the log ratios (Fig. 5) (11).

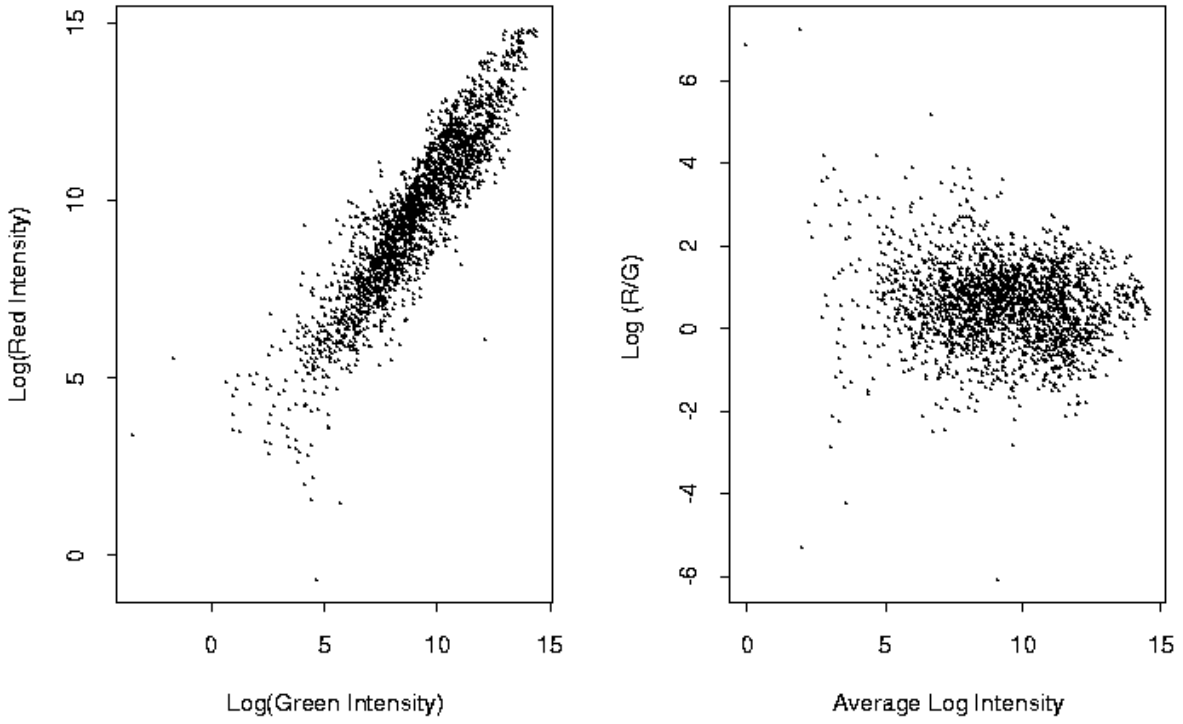

Fig 5. Data displays. When wanting to compare two sets of numbers such as R and G varying over a large range, it is useful to compare $\log_2$ R with $\log_2$ G by plotting their difference $\log_2(R/G)$ against their average $(½)\log_2$ R*G. Doing this we might see something unexpected. By contrast, plotting R against G is typically much less revealing and can give a false unrealistic sense of concordance.
(The illustration was used with permission from Speed, T. http://www.stat.Berkeley.edu/users/terry/zarray/Html/log.html, 2 Nov 2002)

1.2 Normalisation

The purpose of normalisation is to minimize methodological variations in the measured gene expression levels, to display biological differences more clearly and to allow between-slide comparisons. Sources of systematic variation are different labelling efficiencies and dye label concentrations, scanning properties of the dyes and print-tip or spatial effects on the chip. (13-15). Imbalance in the red and green intensities is easily observed when two identical mRNA samples are labelled with different dyes and hybridised on the same slide. In this kind of experiment the red intensities tend to be lower than green intensities and the magnitude of the difference may depend on overall intensity A, resulting in a curvature in the MA-plot (10,14).

1.2.1 Selecting genes for normalisation

Whatever normalisation method used, which set of genes to use has to be chosen:

All the genes on the array
Using all the genes on the chip for normalisation is reasonable when one assumes total gene expression to be approximately equal in reference and sample, *i. e*. only a relatively small fraction of the genes will be significantly differentially expressed. Furthermore one assumes that there is symmetry between up- and down regulated genes, making these changes balance out.

Housekeeping genes
Another approach is to select a subset of genes on the array for normalisation, traditionally so-called housekeeping genes. These genes are believed to have constant expression across a variety of conditions (*e.g.* β-actin). In practice such genes are unfortunately very difficult to identify. It may however be possible to find "temporary" housekeeping genes, i.e. genes with constant expression for particular experimental conditions. A limitation with housekeeping genes is that they tend to be highly expressed, not allowing the estimation of dye-bias when this is an intensity dependant factor (14,15).

Controls
A third alternative is to use spiked controls or a titration series of control sequences. In the spiked controls method, synthetic DNA sequences or DNA sequences from an organism different from the one being studied are spotted on the array and included in the two different mRNA samples at equal amount. On the micrroarray, these spots should have equal red and green intensities and could thus be used for normalisation. (14,15).

1.2.2 Normalisation methods

Normalisation can be based on a number of principles, some of which are better than others. The most widely used methods are:

Global normalisation
Global normalisation assumes that red and green intensities are related by a constant factor. A scaling factor be calculated and used to correct for observed differences, forcing the distribution of the log ratios to have a median zero within each slide:

$$\log_2 R/G \rightarrow \log_2 R/G - c = \log_2 R/(kG) \qquad (1)$$

Usually $c=\log_2 k$ is the median or mean of the log-intensity ratios for the genes. Global median or mean normalisation simply results in a vertical translation of the MA-plot, but does not account for the intensity- and spatially dependent effects often observed(12,14).

Intensity-dependent normalisation
A typical characteristic seen in microarray data is a strong intensity-dependency. Dye bias seems to be dependent on spot intensity, with a greater uncertainty of measurements found at lower intensities. This leads to more unreliable ratios for genes with a low total expression.

Lowess is a robust scatter-plot smoother from the statistical software package R (16), which can perform a local intensity (A) dependent normalisation:

$$\log_2 R/G \rightarrow \log_2 R/G - c(A) = \log_2 R/(k(A)G) \qquad (2)$$

where c(A) is the lowess fit to the MA-plot (14).

Lowess stands for Locally-Weighted Estimation, also known as locally weighted polynomial regression (LWR) and is a method for fitting curves to noisy data by robust locally linear fits. The term "robust" refers to the fact that the polynomial is fit using weighted least squares, giving more weight to points near the point whose response is being estimated and less weight to points further away (17).

Print-tip normalisation
A robotic arrayer typically has 4 by 4 or 2 by 2 print-heads, every grid of spots being printed with the same print-tip. There is always a risk of finding systematic differences between the print-tips, i. e. differences in the length or in the opening of the tips or deformation of some of the tips after long use. The print-tip groups are also potential targets for spatial effects on the slide. A print-tip normalisation applies lowess on each grid separately, considering both spatial and intensity effects:

$$\log_2 R/G \rightarrow \log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G) \qquad (3)$$

where $c_i(A)$ is the lowess fit to the MA-plot for the i:th grid only, $i=1,\ldots,I$ and I represents the number of print-tips (14,15).

When different slides have substantially different spreads in their log-ratios, one can perform an additional *scale normalisation*, enabling comparisons between slide experiments, avoiding one or more slides having undue weight. This might however increase the variability of the log-ratios and should be avoided when differences are fairly small(15).

1.3 Selecting differentially expressed genes

After a proper normalisation procedure of data from several microarray experiments, genes with a significantly changed expression can be found. A problem is that one usually has very few replicates for each gene, but is investigating many genes simultaneously (18). Many data analysis programs sort the genes according to the absolute level of the ratio or $M=\log_2(Cy5/Cy3)$. However this approach risks giving genes with *large* variances a too good chance of being called as differentially expressed. A better alternative is to rank genes according to the value of the t-statistic

$$t = M/(s/\sqrt{n}) \qquad (4)$$

where M is the mean of the $M=\log_2(Cy5/Cy3)$ for any particular gene across a series of n replicate arrays and s is the standard deviation of the M-values. This approach protects against outlier M-values, but is not ideal. Large t-statistic can be driven by an unrealistically *small* value of s, resulting in genes with too small sample variances being called as differentially expressed. A suitable compromise between ranking genes according to ratios and t-statistics is to use a *penalized* t-statistic such as *SAM* or *Bayes* (9).

1.3.1 SAM – Significance Analysis of Microarrays

Tusher *et al* (19) suggests forming a penalized t-statistic $d(i)$

$$d(i) = (x(i)_2 - x(i)_1)/(s(i) + s_0) \qquad (5)$$

where $x(i)_1$ and $x(i)_2$ are the average expression levels for gene (i) in states 1 and 2 respectively and $s(i)$ is the standard deviation of repeated measurements. The penalty, or "fudge factor", s_0 is chosen to minimize the coefficient of variation of d (19).

After computation of d(i) for all genes, permutations of response labels are performed and d(iperm)=average d(i) on permuted data is calculated. $\Delta <$ d(i)-d(iperm) is set by the user and

defines the number of significant genes. Δ is chosen to control the *False Discovery Rate* (FDR) which is the expected proportion of errors amongst the genes selected as significantly differentially expressed (Fig. 6) (20).

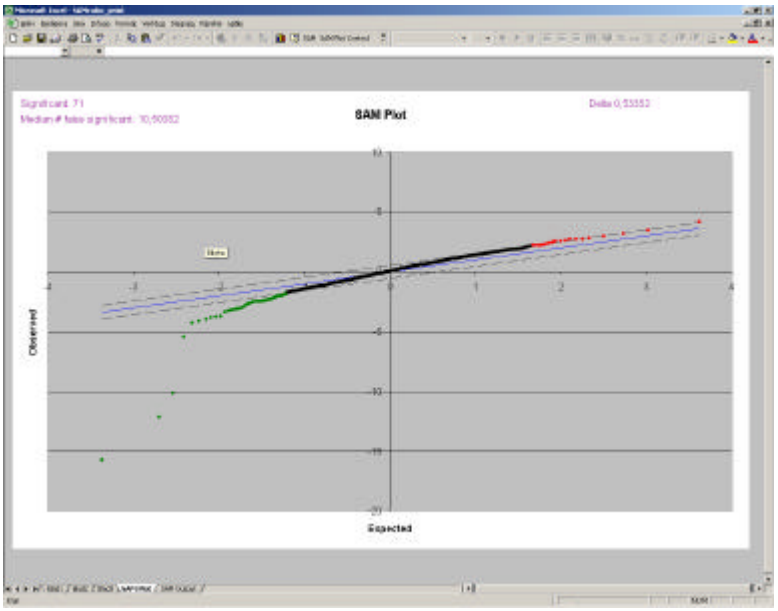

Fig 6. SAM Plot with delta 0.53 and FDR = 10.5/71

Input data has the form of an Excel spreadsheet where the first row has information about the response measurement and all remaining rows have gene expression data, one row per gene. Examples of response formats are *paired data*, with Cy5 and Cy3 intensities or *one class* with logged ratios Cy5/Cy3 (20) (see Table I).

Table I. SAM data Example of paired data response format (a) and one class data (b).

a.

	Cy3[1]	Cy5[1]	Cy3[2]	Cy5[2]
Gene#1	1050	1260	1220	1586
Gene#2	15063	14611	14677	13943
Gene#3	2300	4830	2511	4520

b.

	Log2(Cy5/Cy3)[1]	Log2(Cy5/Cy3)[2]
Gene#1	0.27	0.38
Gene#2	-0.04	-0.07
Gene#3	1.1	0.85

In the case of one class response, the set of expression values for each experiment are multiplied by +1 or –1, with equal probability, i.e. a permutation of signs rather than class labels (20).

1.3.2 Empirical Bayes statistic

An alternative method for finding significant genes is to use an empirical bayesian approach. Lönnstedt and Speed (18) suggests forming a B (from Bayes)-statistic for each gene which is equivalent for the purpose of ranking genes to the penalized t-statistic

$$t=M/\sqrt{(a + s^2)/n} \qquad (6)$$

where *M* is the mean of the $M=\log_2(Cy5/Cy3)$ for any particular gene across a series of *n* replicate arrays, *s* is the standard deviation of the M-values and *a* is a penalty (cf. fudge factor s_0 of SAM, Eq. 5) estimated from s^2.

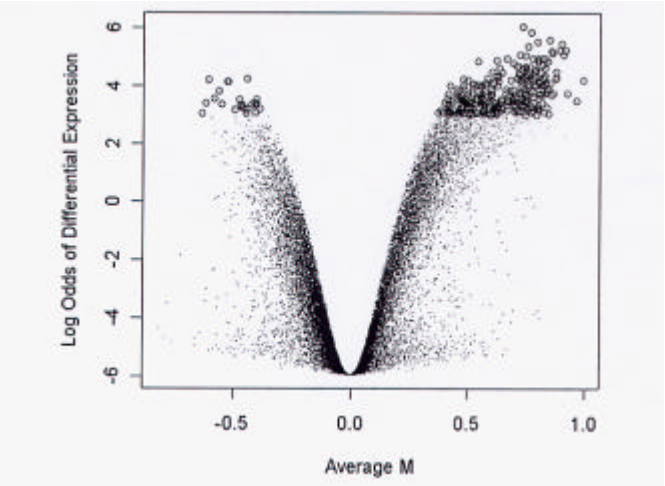

Fig 7. Empirical bayesian approach. Volcano style plot of lodsratio vs M-plot. Genes with *lods* greater than three have been highlighted for follow up and confirmation.

Applying Bayes (Eq. 6) on a set of microarray data, results in a list of Bs, or *lods*, a "Bayes log posterior odds". The list provides a ranking of genes with respect to the posterior probability of each gene to be differentially expressed. It is up to the scientist to choose a suitable cut-off, the number of genes selected depending on the size, aim, background or other factors of the experiment (18).

1.3.4 Own unpublished method: *Eva*

This is a method based on basic probability theory. Looking at the probability of one gene to end up in the top quarter of the gene list on one array by pure chance is ¼. The probability of the same thing happening on two arrays is $1/4^{\wedge 2}$, and so on. To get the overall probability of your genes being at a specific fraction of the gene list by chance, you multiply the probability for one gene by the total number of genes on the chip.

The sensitivity of Eva increases the more replicate arrays you have; the more replicates the larger fraction you can choose without getting unacceptably many genes called by chance. The following example with arrays of 500 genes tries to illustrate this fact:

Table II. Unpublished method "Eva". Example of how #arrays affects the number of genes called by chance

#arrays	3	4	5	3	4	5
up/down definition	±1/3	±1/3	±1/3	±1/5	±1/5	±1/5
up/down by chance	37	12	4	8	1.6	0.32

FDR for this method is defined as #genes called by chance divided by #called genes.

To generate gene lists from Eva, genes are ranked according to ratio and every gene is assigned a number corresponding to its place in the ranking list. The most up-regulated genes are placed in the top and the down-regulated in the bottom of the list. The user defines a percentile of the gene list, which is defined as up- and down respectively. In order to be

counted as up- or down-regulated, the gene has to be found in the specified percentile in lists from *all* arrays.

Genes with to large variance will not appear on the list at all, since it is enough with one gene to drop out of the specified percentile to exclude it from the generated gene list. This could be avoided by for example excluding one array at the time and applying the method on the remaining data. If a gene appears on all lists except one, maybe it is reasonable to include it.

The method Eva could be further refined by completing the ratio-ranking list with a list ranking the genes according to variance, for example by a simple t-statistic. Significantly differentially expressed genes should preferably combine a high ranking on the ratio-lists with a low ditto on the variance ranking-list.

1.4 Genetic adaptations to strength training

Strength training, or more precisely "short bursts of muscle activity against high resistance or by prolonged stretch beyond normal resting length"(21), causes among other effects hypertrophic growth of skeletal muscle. This myofiber hypertrophy is characterized by a general increase in protein constituents of the muscle fibers. The hypertrophic process is partly caused by the cumulative effects of transient changes in gene expression of *specific* genes. The major events underlying muscle growth is however a *general* and non-specific augmentation of protein synthesis within the cells (21)

Most of what is known about hypertrophy is derived from animal studies, much often by studies of cardiac muscle. Hypertrophy may favour a fast-to-slow fiber-type transition associated with shifts in myosin heavy chain (MHC) isoform expression. In contrast to endurance training, hypertrophic growth does not induce expression of mitochondrial enzymes. Changes in the expression of the transcriptional factors c-fos and c-myc, may be part of a cascade leading to (cardiac) cell hypertrophy. Stretch-induced events in skeletal myofibers appear to be similar to the responses of cardiomyocytes, though the specific growth factors involved are probably different, with insulin-like growth factors playing an important role (21).

There are still a very limited number of studies addressing the question of how human skeletal muscle responds to strength training at the molecular level.

2. AIM OF THE PROJECT

The aim were to

i.) perform a quality check of if and how the dye label concentration of Cy3 and Cy5 affects the quality of raw data from a custom made microarray.

ii.) by literature studies find a satisfactory normalisation procedure and apply it on data generated from the chip.

iii.) find alternative statistical methods for identifying differentially changed genes, choosing two methods and compare these with an own unpublished alternative.

iv.) make a biological application by using the data analysis above: gene expression profile was studied in human skeletal muscle before and after a period of strength training.

3. MATERIALS AND METHODS

3.1 cDNA microarrays

A custom made human cDNA microarray Myochip 1.0 from ClonTech (Cat. #CS2003) was utilized. The microarray included 500 selected genes for cell signalling, oxidative stress, angiogenesis, mitochondrial biogenesis, myogenesis, apoptosis, cell cycling and DNA husbandry.

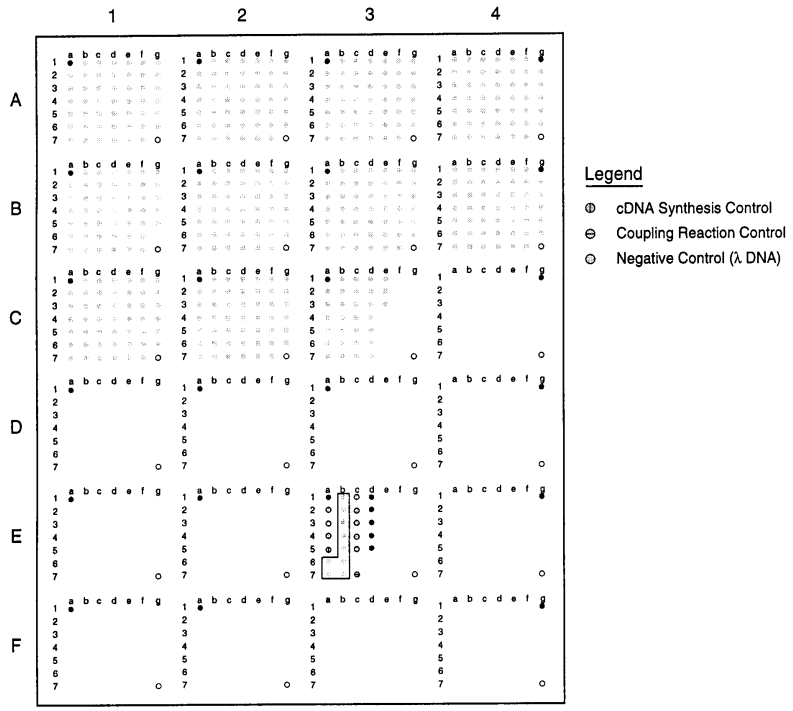

Fig 8. Custom Atlas Glass Microarray (Cat. #CS2003) Myochip 1.0

3.2 Preparation of labelled cDNA

3.2.1 Muscle biopsy and RNA extraction

Percutaneous muscle biopsies from three healthy male subjects, aged 25, were obtained at rest from *m. vastus lateralis*. Total RNA was prepared by the acid phenol method: Biopsies were homogenized 30 sec with a Polytron knife in 1.2 (1 vol.) to 2 ml Denaturating Solution with 0.1 M DTT and 0.5% sarkosyl. 0.1 vol. of 2 M Na acetate (pH 4) was added mixed thoroughly. 1 vol. of H_2O saturated phenol was added and the samples were again mixed well. 0.2 vol. of chloroform iso-amyl-alcohol (49:1) was added, the samples were mixed and incubated for 15 min. at 0-4°C. The samples were centrifuged in a microcentrifuge (4°C) at 10 000g for 30 min. The upper aqueous phase was transferred into new 2 ml tubes. To precipitate, 1 vol. of isopropanol was added and the samples left at -20°C for 30 min. Samples were centrifuged at 10 000g (4°C) for 30 min and the supernatant was removed. The pellets were redissolved in 0.3 initial vol. of DS (w/o DTT and sarkosyl). An equal vol. of isopropanol was added and left at -20°C for 1h. The samples were then centrifuged at 10 000g (4°C) for 30 min. The supernatant was removed and the pellets were washed with 75% ice-cold ethanol. Finally the RNA was dissolved in 50 µl H_2O and stored at –80°C until cDNA synthesis and labelling.

3.2.2 RNA quality and quantification

RNA was quantified spectrophotometrically by absorbance at 260 nm and checked for protein content by examining the A_{260}/A_{280}-ratio. RNA integrity was determined by 1% agarose gel electrophoresis. RNA from the three subjects was pooled and a Bioanalyzer (Agilent Technologies, California, USA) was used to confirm the quality- and quantity check. A_{260}/A_{280}-ratio was above 1.9 and electrophoresis showed intact ribosomal 28S and 18S RNA bands. Bioanalyzer rRNA ratio (28S/18S) was 1.74, well within the reccomended range

3.2.3 cDNA Synthesis and Purification

Indirect labelling was based on a slightly modified version of Atlas™ PowerScript™ Fluorescent Labeling Kit from ClonTech (#:K1860-1). 2.25 µg and 5µg of total RNA was used per labelling reaction, 4 samples of each. For each reaction, a MasterMix of 7.2 µl 5X First Strand Buffer, 3.6 µl 10X dNTP, 3.6 µl DTT, 1.8 µl H_2O, and 1.8 µl PowerScript Reverse Transcriptase was prepared and kept on ice. Because of the relatively low concentration of our RNA, the MasterMix described above is 1.8 times the volume recommended in the manufacturers protocol. 2 µl Random Primer Mix and 1 µl cDNA Synthesis Control was added to each RNA sample, but no extra H_2O needed to be added to reach the proper reaction volume. The samples were heated to 70°C in a PCR thermalcycler for 5 min, cooled to 37°C after which 18 µl MasterMix was added per reaction and and left to incubate at 37°C for 1 hr. The tubes were then incubated at 70°C for 5 min and spinned briefly in a microcentrifuge to collect contents. After cooling tubes to 37°C, 0.2 µl Rnase H was added and the samples were incubated at this temperature for 15 min. Tubes were spinned and 0.5 µl EDTA (pH 8.0) and 2 µl QuickClean resin was added. The samples were vortexed for 1 min. 0.22-µl Spin Filters were inserted into collection tubes, and each sample was transferred into a filter. The tubes were spinned at maximum speed for 1 min. The Spin Filters were removed and 2.2 µl 3M Sodium Acetate and 55 µl ice cold 99% ethanol was added and samples were vortexed. Tubes were placed in -20°C freezer to precipitate the cDNA and then spinned for 20 min in a microcentrifuge (4°C). After pipeting off the supernatant, pellets were washed in 70% ethanol and then dissolved in 10 µl 2X Fluorescent Labeling Buffer.

3.2.4 Fluorescent Dye Coupling

According to the manufacturers recommendations, the labelling kit was supplemented with Cy3- and Cy 5 Mono-Reactive Dye Pack (Amersham Pharmacia Biotech #PA23001 and #PA25001). 5 mM stock solutions of fluorescent dye were prepared by adding 45 µl DMSO to the dye vials, which were then vortexed and spinned. 0.5 µl Coupling Reaction Control Oligo was added to each cDNA sample. 10 µl dye was added to the samples and these were then mixed well and placed at room temperature, wrapped in aluminium foil, for 1 hr. 2 µl 3M Sodium Acetate and 50 µl 99% ethanol was added and the samples placed in a -20ºC freezer for 2 hr to precipitate the labelled target. Samples were spinned for 20 min and the supernatant pipeted off. The pellets were washed in 70% ethanol and dissolved in 100 µl H_2O.

3.2.5 Target Purification

The labelled target was purified using Qiagen´s Quiaquick PCR Purification Kit (#28104) with silica gel spin columns. The protocol was modified according to the manufacturers.

instructions; each wash step was performed a total of three times, using 650 µl of buffer for each wash. In the elution step, elution was performed twice with 30 µl Buffer EB allowing the column to stand for 1 min after adding the buffer. Notice that if the buffer used to bind the DNA to the Qiagen PCR purification columns is not slightly acidic (less than pH 7) the cDNA will bind poorly to the column, resulting in low yields. Also elution efficiency is dependant on pH, maximum elution efficiency being achieved between 7.0 and 8.5.

3.2.6 Dye label measurements

The dye label concentration was determined in a Nanodrop™ ND-1000 Spectrophotometer (NanoDrop Technologies, Inc). Total dye quantity in the 5 µg-samples was 66, 54, 54 and 42 pmol and in the 2.25 µg-samples 18, 18 and 12 pmol. One of the 2.25 µg-samples was below detectable concentration.

3.3 Hybridisation

3.3.1 Hybridisation

Two hybridisations were performed for each RNA quantity, resulting in a total of 4 self-against-self experiments. Microarray slides were hybridised according to the Atlas™ Glass Microarrays User Manual from ClonTech: To yield a final volume of 1.9 ml of target and hybridisation solution, 1.78 ml of GlassHyb Hybridisation Solution per slide was warmed to 50ºC and the labelled targets (2 x 60 µl) were then added. The solution was added to the Hybridisation Chamber and left to hybridise over night at 50ºC.

3.3.2 Washing microarray slides

Washing of the microarrays was performed at room temperature according to the same Atlas™ Glass Microarrays User Manual: The washing procedure was performed in four steps, each step for 10 minutes on an orbital shaker with the Wash Containers in an upright position. Wash 1 was performed in 22 ml GlassHyb Wash Solution (=2X SSC + 1% tween), wash 2a in 2 ml GlassHyb Wash Solution + 20 ml 1X SSC, wash 2b again in 2 ml GlassHyb Wash Solution + 20 ml 1X SSC and wash 3 in 22 ml of 0.1X SSC. Finally the slides were removed from wash 3, and rinsed briefly under running distilled water. Drying of the slides was accomplished through dipping them quickly in iso-propanol and blowing the moisture off the surface with N_2 gas.

3.4 Evaluation of chip quality *vs* dye label concentration

In an attempt to decide how the dye label concentration affected the chip quality, different key properties of the chips were calculated and visualised in an Excel-spreadsheet. More specific, the slides were examined with respect to: spots with intensities below 2- and 4 times background intensity, saturated spots, flags, spot intensity expressed as mean-, median- and 25- and 75 percentile intensity, background intensity expressed as mean and median. The median spot intensity was also compared to overall median background intensity as well as to its local background.

3.5 Data analysis

3.5.1 Scanning

The hybridised slides were scanned with the GenePix® 4000B scanner (Axon Instruments, California, USA). PMT settings were chosen to balance the two channels, using the entire dynamic range (0-65535), but trying to avoid saturation. An image of each of the Cy3 and Cy5 channels was generated and saved as a 16 bit TIFF file.

3.5.2 Image extraction and data analysis

GenePix Pro 3.0 Microarray Analysis Software (Axon Instruments, California, USA) was utilized to extract data from the TIFF files. A grid pattern was placed on the image to mark the location of the spots. Subjectively judged bad quality-spots were manually marked with a "flag".

The GenePix Pro software measures the spot intensity of Cy3 and Cy5 and additionally computes the local background around each spot. The resulting calculations were saved as an Excel-type spreadsheet for further analyses.

The Cy5/Cy3-ratio was calculated from median pixel intensities of the spot. No background subtraction was used, as it has shown to increase variation (22).

3.5.3 Normalisation

The raw data was visualised in a MA-plot as described in the background section. The statistics language R together with the package sma (Statistics for Microarray Analysis) were used to perform various forms of normalisations and data analysis (16).
The normalised M and A values were exported to an output file that can be opened in *e.g.* Excel.

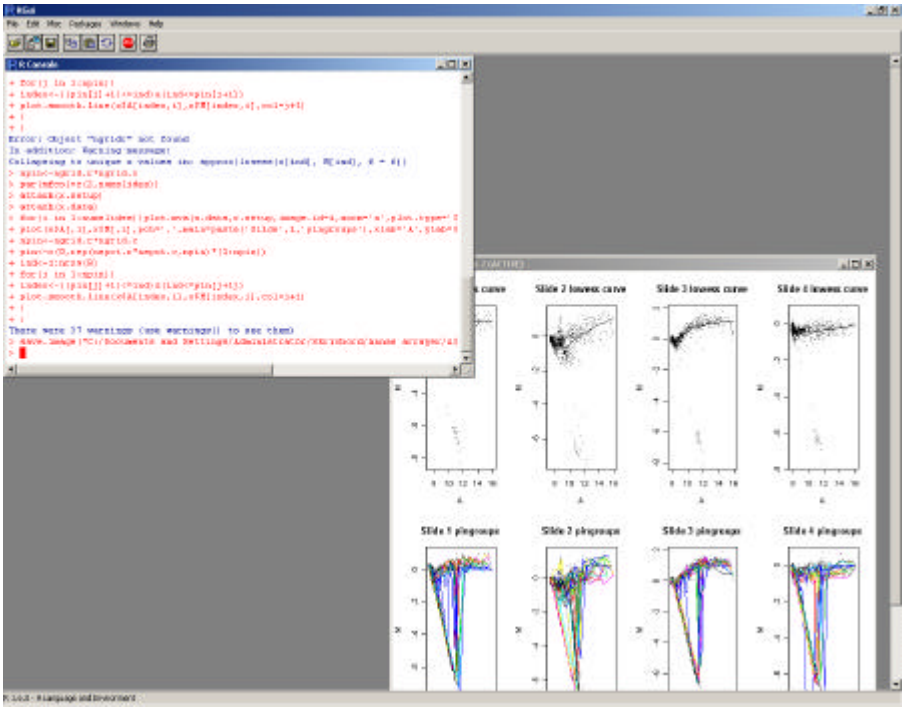

Fig 9. Data analysis The R user interface (version 1.3.1)

3.5.4 Statistics

The normalised data was examined using three different statistical methods; *"Eva"*, *SAM* (Significance Analysis of Microarrays) and *Bayes*. The SAM analysis was carried out using the Excel Add-in *SAM*, Bayes was carried out in R and Eva was performed "manually" in Excel.

In order to compare the outcome and look for concordance between the various statistical methods, a "FDR" for Eva's method was calculated based on 1/3 and 1/5 chosen as up/down-limits by dividing #genes called by chance with #called genes. Gene lists from SAM was generated using the same FDR. As there is no evident way of calculating a FDR for the Bayes method, the mean number of the sum of up- and down regulated genes in Eva's and SAM approach was used for comparison.

Normalised intensities were calculated from M and A values according to the definition of M and A:

$M = \log_2 Cy5/Cy3$
$A = \log_2 \sqrt{(Cy5*Cy3)}$
$\Rightarrow$ Cy5=2^(A+0.5*M), Cy3=2^(A-0.5*M),

The SAM gene lists were compared to Bayes and Eva.

3.6 Biological application of data analysis procedure

Six healthy individuals performed weight lifting with emphasis on leg training, three times a week, under supervision of a personal trainer, for a three weeks period. Needle muscle biopsies were taken from *m. vastus lateralis* 12 hours before and 12 hours after the last training session. Total RNA was prepared as described above. Preparation of labelled cDNA followed the same procedure as described above, using 5 µg of total RNA per reaction. The biopsy taken before training was labelled with Cy3 and the after training biopsy was labelled with Cy5. Hybridisation of the chips and data analysis followed the procedure described above.

4. RESULTS

4.1 Evaluation of chip quality vs. dye label concentration

Row 3 in Table III a. and b. show that the same amount of total RNA resulted in somewhat different dye label quantities; 2.25 µg in the range of 12-18 pmol, 5 µg in the range of 42-66 pmol. Moreover there was no systematic difference between the incorporation efficiency of the two fluorophores during indirect cDNA labelling.

Row 6 and 7 show that increasing the amount of RNA from 2.25µg to 5µg, makes the fraction of spots with intensities below 2 times the background in the red and green channel decrease from 45% to 36% and spots below 4 times the background decrease from 60% to 56%. Consequently, using more RNA gives fewer (-50)-flags, i.e. spots too weak to be found by GenePix. The number of spots over saturation is primarily affected by the skills of the person setting PMT voltage, and cannot be related to dye label concentration.

The tables also show that doubling the RNA gives about a 1.3 increase in median spot intensity. The corresponding increases for the 25:th and 75:th percentile spots are 1.3 and 1.5 respectively. The median background intensity is affected by the RNA quantity by approximately the same factor as the median spot intensity, 1.2. Removing the Cy3 background from chip2, which has a small green fluorescent stain, gives an increase of an even more similar magnitude (1.25). The median spot intensity-median background ratio was 14% higher in the 5µg-chips.

Table III. A display of different key properties of the four self-against-self hybridisations, carried out using different amounts of total RNA per labelling reaction. *PMT* gives information about scanner settings. *Norm.factor:RatioOfMed* is the numerical constant the ratios (of median spot intensities) should be divided with in a global normalisation. *# spots>65535* tells us whether PMT was set too high. Spots are flagged *(-50)* when GenePix fails to find them, while sub-standard spots *are flagged (-100). Median spot* is the median spot intensity. *Spot 25:th* -and *75:th percentile* give the intensity of these spots, which together with Median spot are trying to display the intensity distribution. *Background median spot* is the local background of the median spot. *Median background* is the median of all local spot backgrounds on the chip.

a. shows data generated from chip 1 and chip 2, with 2.25 µg of total RNA per reaction.

1.Total RNA	**2.25 ug**				***Mean***
2.	**1 Cy5**	**1Cy3**	**2 Cy5**	**2 Cy3**	
3. Dye label quant. [pmol]	18	18	?	12	
4. PMT	740	690	750	700	
5. Norm.factor:RatioOfMed	0,91		1,06		
6. spots<2xbackground	206	223	257	257	*236*
7. spots<4xbackground	285	302	319	354	*315*
8. spots>65535(=maxint)	5	3	7	1	
9. (-50)-flags	99		93		
10. (-100)-flags	2		71		
11.Median spot	677	701	613	841	*708*
12. Spot 25:th percentile	289	392	321	437	*360*
13. Spot 75:th percentile	3249	2809	2257	2314	*2657*
14. Backgr. median spot	206	419	360	261	*312*
15. Median background	211	255	223	295	*246*
16. Mean background	223	289	288	428	307
17.Medspot/Backg. medspt	3,29	1,67	1,70	3,22	*2,47*
18.Medspot/Med backgr	3,21	2,75	2,75	2,85	*2,89*

b. shows chip 3 and chip 4, with 5µg of total RNA.

1. Total RNA	**5 ug**				***Mean***
2.	**3Cy5**	**3 Cy3**	**4 Cy5**	**4 Cy3**	
3. Dye label quant. [pmol]	54	54	42	66	
4. PMT	750	700	770	720	
5. Norm.factor:RatioOfMed	0,68		1,34		
6. spots<2xbackground	166	205	199	191	*190*
7. spots<4xbackground	258	315	299	293	*291*
8. spots>65535(=maxint)	25	4	8	9	
9. (-50)-flags	25		63		
10. (-100)-flags	4		1		
11. Median spot	1093	806	811	1047	*939*
12. Spot 25:th percentile	454	477	389	538	*465*
13. Spot 75:th percentile	5746	2963	3073	3992	*3944*
14. Backgr. median spot	260	289	283	408	*310*
15. Median background	264	290	259	334	*287*
16. Mean background	273	295	261	341	*293*
17. Medspot/Backg.medspt	4,20	2,79	2,86	2,61	*3,12*
18. Medspot/Med backgr.	4,14	2,78	3,13	3,13	*3,30*

4.2 Normalisation

Cy3/Cy5-scatter plots of extracted intensity raw data from the four chips (Fig 10) showed that the balance between the red and the green channel varied between the chips, indicating the need for a global normalisation, setting the median ratio to zero. Further on, you could see the importance of using appropriate PMT settings to avoid spot saturation, not to underestimate ratios.

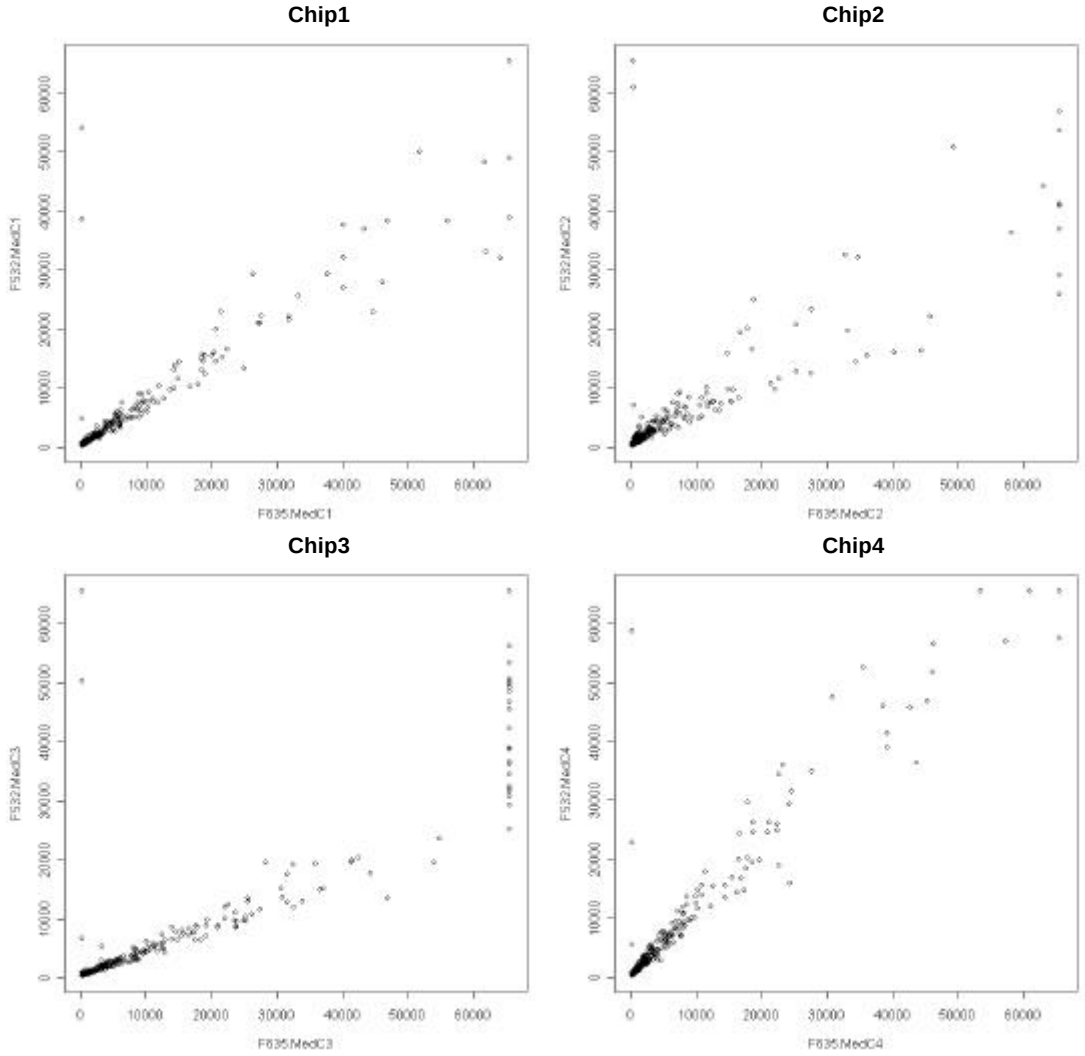

Fig 10. Cy3/Cy5 scatter plots of raw data for the 4 self-against-self hybridisations. In chip3 more spots have reached saturation than in the other chips.

When data was further analysed in R (Fig 11), a strong intensity dependency of ratios was observed in MA-plots of raw data, giving low values of log ratio M for low mean log dye intensity A spots. Data was plotted in different steps of the normalisation procedure; exploring raw-data, lowess normalised data and print-tip lowess normalised data. Based on observation of the MA-plots, a within-print tip group lowess normalisation was performed. This normalisation seemed to eliminate the intensity-dependancy of A, fitting data around log-ratios of zero.

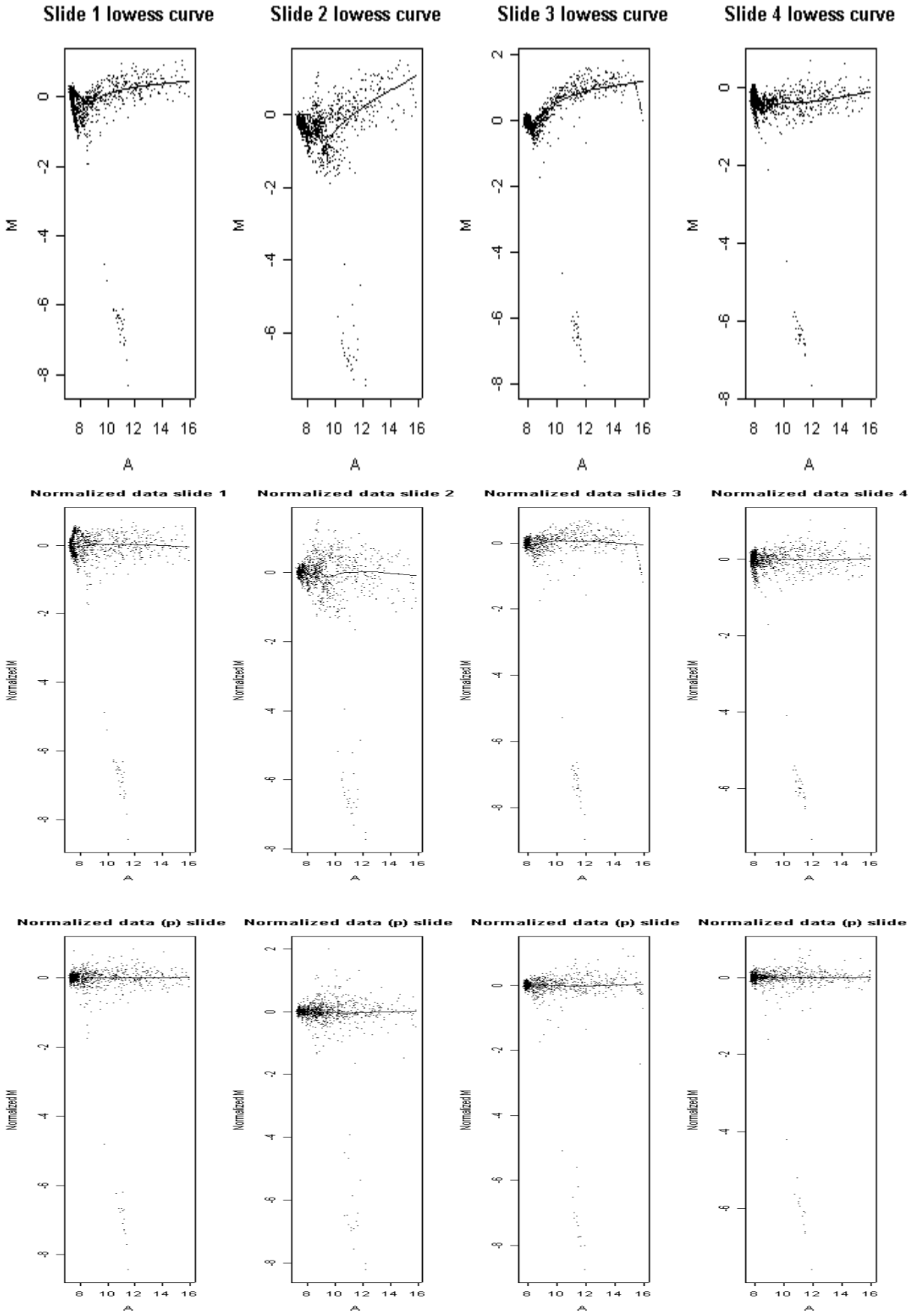

Fig 11. **MA-plots** of data from the 4 self-against-self hybridisations, showing the average M-value vs. the average mean log dye intensity for each gene. Different steps in the normalisation procedure: The upper row displaying raw data, the middle row is data after lowess normalisation and the last row shows data after print-tip lowess normalisation. Notice how the intensity dependency of A seems to be eliminated after the print-tip lowess normalisation, the data being fitted around log-ratios of zero.

Histograms over log ratios confirmed a satisfactory normalisation procedure, with ratios centered around zero after print-tip lowess normalisation (Fig 12) . Histograms also showed that no slide differed substantially in its spread of log-ratios, making a scale normalisation redundant.

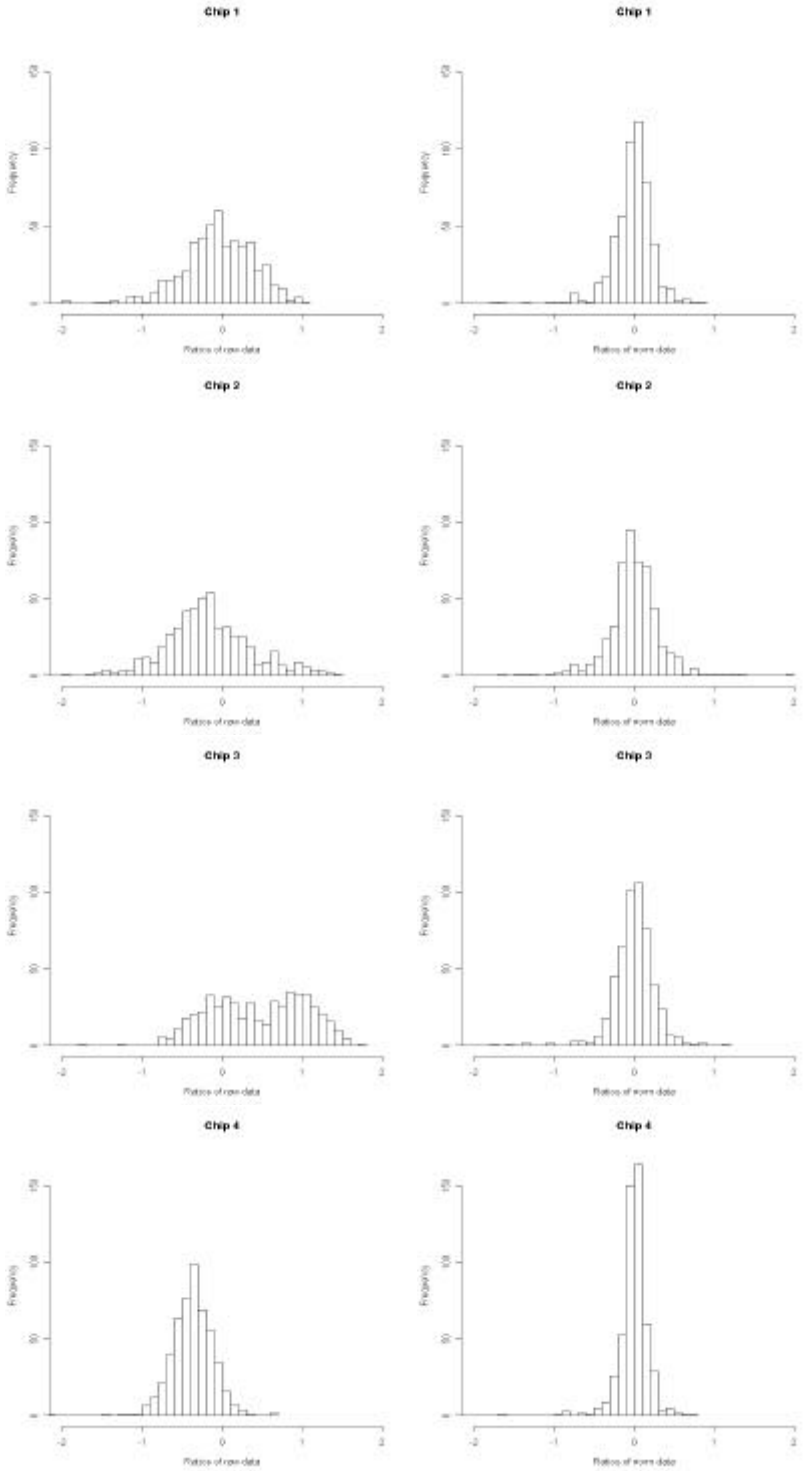

Fig 12. **Histograms** showing ratios log2(Cy5/Cy3) before (left) and after (right) print-tip lowess normalisation, chip 1-4.

4.3 Statistics

Using Eva's method with 1/3 and 1/5 chosen as up/down-limits resulted in two lists. The narrower limit of 1/5 resulted in 16 genes called as up- and 22 genes called as down regulated. Setting the limit to 1/3 generated 48 genes on the up side and 40 on the down side. Defining FDR as #genes called by chance divided by #called genes gave the following two FDRs:

#genes called by chance $_{(1/5)}$ =2 x (524/5∧4)= 1.68
#called genes$_{(1/5)}$= 16+22 = 38
⇒ FDR$_{(1/5)}$=13/88=0.044, *i.e. 4.4%*

#genes called by chance $_{(1/3)}$ =2 x (524/3∧4)=13
#called genes$_{(1/3)}$= 48+40 = 88
⇒ FDR$_{(1/3)}$=13/88=0.147, *i.e. 14.7%*

Column 8 in Table IV. show that 76% of the genes called by SAM, run with Cy3- and Cy3 intensities as response format ("SAM int."), are also called by Eva and Bayes at FDR 4.4%. At a higher FDR, only 61-65% of the SAM int.-genes are found by Eva and Bayes. Except for Eva at FDR 14.7%, more genes are consistently found to be down- than up-regulated, with the greatest imbalance between up and down for SAM int. at FDR 4.4%.

To examine whether the intensity dependency of log ratios was eliminated, the median intensity (of the mean intensity of the four chips) of the top ten most up- and down regulated genes was calculated. Looking at the genes called by SAM int., gave a median intensity of 9.29 for the down-regulated genes, and 11.14 for the up-regulated. These figures can be compared to the overall median spot intensity of 9.66.

Table IV. Statistics Comparison of three different statistical methods (column 1), using FDR 4.4 and 14.7% respectively (column 2). Column 3-4 gives the number of genes being called as up- and down regulated, with the total number of changed genes in column 5. As Bayes approach only ranks the genes with respect to probability of being differentially expressed, the number of called genes was defined as the mean number of the sum of up- and down regulated genes in Eva's and SAM int. approach. The remaining columns show how well the methods agree with SAM int. (highlighted squares), column 6-8 giving the absolute number and column 9 the percentage of SAM int. genes showing up on the lists of the other methods.

1.	2. FDR%	3. Up	4. Down	5. Total genes	6. up fr SAM int.	7. down fr. SAM int.	8.Total f. SAM int.	9. % genes fr SAM int
SAM int.	4.4	2	15	17	x	x	x	x
SAM int.	14.7	44	61	105	x	x	x	x
Eva	4.4	16	22	38	1	12	13	76
Eva	14.7	48	40	88	26	38	64	61
Bayes	x	x	x	28	1	12	13	76
Bayes	x	x	x	97	25	43	68	65

4.4 Biological application: significantly changed genes

Table V. Biological application Genes, identified by SAM and Bayes, with significantly changed expression in six individuals who performed strength training for three weeks.

Up-regulated genes	Function	Cy5/Cy3	M
adrenergic, alpha-2B-,receptorSHC (Src homology 2 domain-containing) transforming protein 1	enzyme inhibitor	1.33	0.41
desmin	muscle structure	1.32	0.40
thioredoxin	redox reactions	1.20	0.26
serum response factor (c-fos serum response element-binding transcr. factor)	transcription	1.21	0.27
peroxisome proliferative activated receptor, alpha	transcription	1.22	0.29
general transcription factor IIH, polypeptide 2 (44kD subunit)	transcription	1.20	0.26
xeroderma pigmentosum, complementation group A	DNA repair	1.78	0.83
Down-regulated genes			
cyclin A2	cell cycle regulator	0.48	-1.06
uncoupling protein 2 (mitichondrial, proton carrier)	energy-intermediate enzyme	0.72	-0.48
cyclin B1	cell cycle regulator	0.74	-0.44
MADS box transcription enhancer factor 2, polypeptide A (myocyte enhancer factor 2A)	transcription	0.81	-0.31
polymerase (DNA directed), delta 2, regulatory subunit (50 kD)	replication	0.80	-0.32
heat shock 70 kD protein 1B	protein assembly	0.52	-0.95
thymine-DNA glycosylase	DNA repair	0.79	-0.33
early growth response 1	transcription	0.76	-0.40
integrin, alpha 1	collagen receptor	0.79	-0.34
growth differentiation factor 8	muscle growth (neg. reg.)	0.82	-0.29
growth arrest and DNA-damage-inducible, alpha	DNA repair	0.84	-0.25

Running biological data in SAM And Bayes resulted in gene lists where quite a few of the top genes in the lists from the self-against-self hybridisation appeared. Using the self-against-self hybridisation-lists as a filter, removing these genes from the biological data lists, resulted in the list above. The genes in Table V. were identified by both SAM and Bayes as differentially expressed at a FDR of 14.7%, with the down-regulated genes consistently assigned higher statistic scores.

The down-regulated genes have ratios, in the magnitude of ~0.5 to 0.85, up-regulated ditto have ratios of ~1.2-1.8. Up-regulated genes were related to the activity of transcription, enzyme inhibition, muscle structure and DNA-repair. An attenuated activity was found from genes related to the groups of cell cycle, energy metabolism, transcription, replication, DNA-repair etc.

In the case of the biological data, looking at the median intensity $\log_2$(Cy5*Cy3) of the top ten positive and negative genes called by SAM, gave values of 10.11 (9.4-11.1) for the ten most negative and 11.35 (9.9-12.6) for the top ten positive. The overall median of mean spot intensities was 10.46. Corresponding values in unlogged format were 1100 (680-2200), 2600 (960-6200) and 1400 respectively.

5. DISCUSSION

This investigation shows that dye label incorporation in the magnitude of 15 pmol per dye, results in chip data where 55% of the spots have both red and green channel intensities 2 times higher than background. An increase of dye label quantities to approximately 50 pmol, make the proportion of spots having a signal-to-background ratio greater than 2, increase to 65%. These figures are interesting with respect to how you should interpret information from the Nanodrop spectrophotometer, deciding if your labelling is satisfactory enough to allow a subsequent hybridisation. There are no earlier published studies relating dye label quantity to spot intensities, but reading protocols from various microarray labs give recommendations of dye incorporation per sample ranging from >200 pmol (23), 30-60 pmol (24) and most consistent with this study; optimal range 20-50 pmol, minimal 15 pmol (25). Looking at our data, minimal dye incorporation recommendation could be set to 10 pmol.

MA-plots (see Fig. 6) of raw data showed a strong intensity-dependency of ratios, an intensity-dependency, which was most effectively eliminated by a print-tip lowess normalisation. This normalisation method is well supported in literature (2, 9-15, 18). What might contradict a total success of this print-tip lowess normalisation, is that median spot intensity of the ten most up-regulated genes is slightly higher than the total median spot intensity, while the median spot intensity of the ten most down-regulated genes is slightly lower. This means that there may be an overestimation of down- and up-regulated genes.

The statistical methods SAM and Bayes, for finding differentially expressed genes showed a good concordance, especially at low FDRs. This agrees with both literature (9,18,26,27) and the fact that the two methods are based on the same principle of a penalized t-statistic. Despite the fact that our own unpublished method "Eva" is based on completely different principles than SAM and Bayes, the concordance between this method and SAM was almost as good as the agreement between SAM and Bayes. The principal advantage of Eva is perhaps its simplicity and straightforwardness. Rather than replacing some of the very sophisticated statistical methods, it can serve as a useful tool to get a fast grip of what your data in indicating.

Common to all three statistical methods is that genes with too large variance will drop, or not appear at all in the gene lists. If you are unlucky, large variance is caused by bad spot quality on one array. To avoid missing such a gene, an approach is to exclude data from one (or more) array at the time, applying the method on the remaining data. If a gene appears on all lists except one, maybe it is worth taking a closer look at.

It is hard to give a satisfactory explanation to why we found genes with significantly changed expression in this self-against-self experiment, except for a certain number that would be expected to appear by chance. If there were something wrong with the spots of these genes, it would be reasonable to look more carefully at them if they appeared as top candidates in a biological experiment. If they tend to bind more strongly to one of the colours Cy3 or Cy5, a *dye-swap filter, i. e.* to use two hybridisations for two mRNA samples, with dye-assignment reversed in the second hybridisation, would hopefully remove this effect.

This study focuses on methodological aspects of microarrays and microarray data. In addition an application of the data analysis methods on a biological experiment was performed: six individuals performed strength training for a three weeks period. The identified genes represent several functional classes and it is hard to see a specific pattern. Despite the fact that

hypertrophic growth is not expected to affect expression of mitochondrial proteins (21), we found down-regulation of two genes encoding mitochondrial proteins; uncoupling protein 2 and heat shock 70 kD protein 1B. Confirming literature reporting that changes in the expression of c-fos induces hypertrophy (21), we found up-regulation of a c-fos serum response element-binding transcription factor.

Previous work characterizing gene expression profile in human skeletal muscle after strength training is limited. Roth *et al.* (28) used a cDNA microarray representing 4,000 human genes, and managed to identify 69 genes as differentially expressed (>1.7-fold) in response to strength training. None of these genes do however correspond to our findings. A similarity between our data and Roth's findings, is that a majority of the most differentially expressed genes were down-regulated. One might however question the reliability of Roth´s data, since he does not perform an adequate data analysis with respect to normalisation and has no statistical tools for selecting the genes. Another difference between our and Roth´s study was the length of the training study: The study of Roth examined the effects of strength-training for 9 weeks, while our study was a short-term training study of 3 weeks.

The identified genes showed quite modest ratios, maybe because of a too weak stimulus or because of the general nature of the major events underlying muscle growth (21). Another problem in skeletal muscle array experiments is the reported high interindividual variability, which can obscure general patterns of expression (28,29).

5.1 Conclusions

The custom made microarray Myochip 1.0 provides a validated tool for the study of gene expression in human muscle biopsy material. To assure quality of raw data, dye-label concentration should be measured (i.e. in the Nanodrop spectrophotometer), minimal 10 pmol CyDye incorporation per sample is recommended. Normalisation of raw data is necessary to correct for systematic variation of ratios, print-tip group lowess normalisation appearing as a good choice. After normalisation, two well-working statistical methods to determine differentially expressed genes are SAM and Bayes, both showing high concordance with our own alternative "Eva". The data procedure has been tested and works well on biological data, predicting significant gene expression profile changes in response to strength training.

ACKNOWLEDGEMENTS

First of all I would like to thank my supervisor Prof. Eva Jansson for giving me the opportunity to work at the Department of Clinical Physiology at Karolinska Institutet, also for taking her time to endless discussions and always believing in me. I would like to thank Kristina Miras de Gea for being such a support in the lab. I also own a great gratitude to everybody at the WCN Microarray Platform in Uppsala, for letting me participate in their course on analysis of microarray data. Special thanks is addressed to Malin Larsson and Mårtin Fryknäs for hours of phone support. I am very grateful to Elsebrit Ljungström at KI Chip core facility for letting us use her equipment and giving valuable advice on the laboratory work. Thanks Lina Goldschmidt for letting us use your lab and giving me the chance to analyse your data. Thanks Anna, Barbara, Maria, Mona, Jessica, Helene and Thomas for great support and making my time at Huddinge so pleasant. Finally, I would like to thank my examiner Ass Prof. Carl-Johan Sundberg.

REFERENCES

1. Klug WS and Cummings MR. The Genetic Code and Transcription. In: Concepts of Genetics. (6th ed). New Jersey: Prentice Hall, Inc., 2000: chapt. 13, 349-79.
2. Holloway A J *et al*. Options available – from start to finish – for obtaining data from microarrays II, Nature genetics supplement 2002;32: 481-89.
3. Atlas ™ Glass Microarrays User Manual. 25 July 2001.
4. Amersham biosciences: Trouble-shooting microarray experiments. 2002.
5. ClonTech Laboratories, Inc. http://www.clontech.com/archive/JAN02/Powerscript.shtml . (1 Sep. 2002).
6. Yu J *et al*. Technical brief: Evaluation and optimisation of procedures for target labelling and hybridisation of cDNA microarrays, Molecular Vision 2002;8:130-7.
7. NanoDrop Technologies, Inc. http://www.nanodrop.com/microarrays.shtml . (1 Sep. 2002)
8. Axon Instruments, Inc. GenePix™ Pro User's Guide, Apr 2000.
9. Smyth GK *et al*. Statistical issues in cDNA Microarray Analysis. Functional Genomics: Methods and Protocols, accepted 2002.
10. Dudoit S *et al*. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Technical report #578. 2000.
11. Speed T. http://www.stat.Berkeley.edu/users/terry/zarray/Html/log.html . (1 Sep. 2002).
12. Quackenbush J. Microarray data normalization and transformation, Nature genetics supplement 2002; 32: 496-501.
13. Bilban M *et al*. Normalizing DNA Microarray Data. Curr Issues Mol Biol 2002;4: 57-64.
14. Yang YH *et al*. Normalization for cDNA Microarray Data: a robust composite method addressing single and multiple slide systematic variation, Nucleic Acids Research 2002;30 (4): 1-10.
15. Yang YH *et al*. Normalization for cDNA Microarray Data, SPIE BiOS, San Jose 2002.
16. Cran R-project. http://cran.r-project.org/src/contrib/PACKAGES.html#sma (15 Jun. 2002).
17. Engineering Statistics Handbook. http://www.itl.nist.gov/div898/handbook/index.htm. (2 Nov. 2002).
18. Lönnstedt I and Speed T. Replicated microarray data, Statistical Sinica, accepted 2002.
19. Tusher *et al*. Significance analysis of microarrays applied to the ionising radiation response. Proc Natl Acad Sci USA 2001;98(9):5116-21.
20. Chu G *et al*. SAM – "Significance Analysis of Microarrays" Users guide and technical document. 2002.
21. Williams RS and Neufer PD.Regulation of gene expression in skeletal muscle by contractile activity. In: Handbook of Physiology. Exercise: Regulation and Integration of Multiple Systems. Betheseda, MD: Am Physiol Soc, 1996: sect 12, part 3, chapt. 25, 1024-1145.
22. Manduchi E *et al*. Comparison of different labelling methods for two-channel high-density microarray experiments. Physiol Genomics 2002;10:169-179.
23. TIGR SOP. http://atarrays.tigr.org/PDF_Folder/Aminoallyl.pdf (4 Jan. 2003).
24. Biogem. http://microarrays.ucsd.edu/protocols/probequant.php (4 Jan. 2003).
25. Biotechnology Research Institute www.bri.nrc.ca/microarraylab/micro/protocolscDNA_e.html (4 Jan. 2003).

26. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 2002;18(4):546-554.
27. Broberg P. Ranking genes with respect to differential expression. Genome Biology 2002;3(9):preprint0007.1-0007.23.
28. Roth SM *et al*. Influence of age, sex, and strength training on human muscle gene expression determined by microarray. Physiol Genomics, 2002;10:181-190.
29. Bakay M. Sources of variability and effect of experimental approach on expression profiling data interpretation. BMC Bioinformatics, 2002;3:4.

APPENDIX I.

Differentially expressed genes in the four self-against-self hybridisations

Top 50 list of genes called as differentially expressed by both SAM and Bayes at FDR 14.7%, in the order they were called by Bayes statistic.

Gene name	ClonTech ID	Ratio
neutral sphingomyelinase (N-Smase) active	A1a2	0.63
gap junction protein, alpha4, 37 kD	B4g2	0.60
protein kinase, AMP-activated, gamma 2 n	C3a2	0.70
polymerase (DNA directed), delta 1, cata	A3a2	0.64
tight junction protein 1 (zona occludens)	C1a2	0.55
X-ray repair complementing defective	A4g2	0.53
TATA box binding protein (TBP)-associate	A3f1	1.08
peroxisome proliferative activated receptor	C3c1	0.97
creatine kinase brain	B1d7	1.27
Janus kinase 3 (a protein tyrosine kinase)	B2e3	1.16
aldehyde dehydrogenase 2, mitochondrial	C3a6	0.80
tumour necrosis factor superfamily	C1a3	0.83
heat shock 27kD protein 2	B2a2	0.65
ubiquitin C	B2c7	1.11
TRAF and TNF receptor associated protein	C1b1	0.97
transcription factor 4	C1c2	1.07
TATA box binding protein (TBP)-associate	A3e6	1.46
proliferating cell nuclear antigen	B1a5	0.96
histone deacetylase 5	B1a3	0.95
natriuretic peptide precursor B	C3a3	0.99
myosin, heavy polypeptide 1, skeletal muscle	C2a2	0.94
tumour protein p53 (Li-Fraumeni syndrome)	A3d5	0.99
ubiquitin protein ligase E3A (human papi	B4a2	0.99
estrogen-related receptor gamma	C2b6	0.95
microsomal glutathione S-transferase 1	B4b6	0.98
glucagon	B2b5	0.96
gap junction protein, alpha 5, 40kD (con	B4g4	0.91
cellular retinoic acid-binding protein 2	B2c2	1.13
fibroblast growth factor receptor 4	A1d7	0.99
nuclear cap binding protein subunit 1, 8	A4f7	0.99
postmeiotic segregation increased (S. ce	A3d6	0.99
ribosomal protein S9	B2c3	1.06
tumor necrosis factor receptor superfami	C1b2	0.99
nuclear factor of activated T-cells, cyt	A2d2	1.00
ELK1, member of ETS oncogene family	B3c5	0.99
cAMP responsive element binding protein	C2a5	0.97
protein phosphatase 3 (formerly 2B), cat	B3a3	0.81
uncoupling protein 3 (mitochondrial, pro	B1e4	0.99
neuropeptide Y receptor Y1	C2d6	1.50
tyrosine 3-monooxygenase/tryptophan 5-mo	A1b1	0.92
protein kinase, AMP-activated, beta 1 no	B1c2	0.94
adrenergic, alpha-1B-, receptor	C2d5	0.98
nuclear respiratory factor 1	C3c3	1.10
retinoic acid receptor, alpha	B2c1	0.99
protein disulfide isomerase	B4a6	0.91
glutathione peroxidase 1	B4b2	1.07
glucocorticoid modulatory element bindin	B2a7	1.26
nitric oxide synthase 1 (neuronal)	C2a4	0.83
ubiquitin	E3b2	1.36
unactive progesterone receptor, 23 kD	B2b2	0.96